EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

Research in Computing Science

Vol. 155 No. 5
May 2026

Research in Computing Science

Series Editorial Board

Editors-in-Chief:

Grigori Sidorov, CIC-IPN, Mexico
Gerhard X. Ritter, University of Florida, USA
Jean Serra, Ecole des Mines de Paris, France
Ulises Cortés, UPC, Barcelona, Spain

Associate Editors:

Jesús Angulo, Ecole des Mines de Paris, France
Jihad El-Sana, Ben-Gurion Univ. of the Negev, Israel
Alexander Gelbukh, CIC-IPN, Mexico
Ioannis Kakadiaris, University of Houston, USA
Petros Maragos, Nat. Tech. Univ. of Athens, Greece
Julian Padget, University of Bath, UK
Mateo Valero, UPC, Barcelona, Spain
Olga Kolesnikova, ESCOM-IPN, Mexico
Rafael Guzmán, Univ. of Guanajuato, Mexico
Juan Manuel Torres Moreno, U. of Avignon, France
Miguel González-Mendoza, ITESM, Mexico

Editorial Coordination:

Alejandra Ramos Porras

Research in Computing Science, Año 25, Volumen 155, No. 5, mayo de 2026, es una publicación mensual, editada por el Instituto Politécnico Nacional, a través del Centro de Investigación en Computación. Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, Ciudad de México, Tel. 57 29 60 00, ext. 56571. https://www.rcs.cic.ipn.mx. Editor responsable: Dr. Grigori Sidorov. Reserva de Derechos al Uso Exclusivo del Título No. 04-2026-043011360400-102. ISSN: en trámite, ambos otorgados por el Instituto Politécnico Nacional de Derecho de Autor. Responsable de la última actualización de este número: el Centro de Investigación en Computación, Dr. Grigori Sidorov, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othon de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738. Fecha de última modificación 01 de mayo de 2026.

Las opiniones expresadas por los autores no necesariamente reflejan la postura del editor de la publicación.

Queda estrictamente prohibida la reproducción total o parcial de los contenidos e imágenes de la publicación sin previa autorización del Instituto Politécnico Nacional.

Research in Computing Science, year 25, Volume 155, No. 5, May 2026, is published monthly by the Center for Computing Research of IPN.

The opinions expressed by the authors does not necessarily reflect the editor's posture.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording or otherwise, without prior permission of Centre for Computing Research of the IPN.

Volume 155(5)

Advances in Artificial Intelligence

León Palafox (ed.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2026

ISSN: in process

Copyright © Instituto Politécnico Nacional 2026
Formerly ISSNs: 1870-4069, 1665-9899

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.rcs.cic.ipn.mx
http://www.ipn.mx
http://www.cic.ipn.mx

The editors and the publisher of this journal have made their best effort in preparing this special issue, but make no warranty of any kind, expressed or implied, with regard to the information contained in this volume.

All rights reserved. No part of this publication may be reproduced, stored on a retrieval system or transmitted, in any form or by any means, including electronic, mechanical, photocopying, recording, or otherwise, without prior permission of the Instituto Politécnico Nacional, except for personal or classroom use provided that copies bear the full citation notice provided on the first page of each paper.

Indexed in LATINDEX, DBLP and Periodica

Electronic edition

ISSN 1870-4069

Table of Contents

Page

ISSN 1870-4069

Trading Algorítmico Ético (TAE): El valor de la capa ética para un sistema de inteligencia artificial

Vladimir Salazar Altamirano

Universidad Panamericana, Facultad de Ingeniería,
México

0252444@up.edu.mx

Resumen. El trading algorítmico ha transformado significativamente los mercados financieros al permitir decisiones rápidas basadas en modelos predictivos de aprendizaje automático. Sin embargo, es común que estos sistemas carezcan de transparencia y explicabilidad, lo que dificulta la comprensión por parte de los usuarios sobre cómo y por qué se toman ciertas decisiones automatizadas. Este desafío se intensifica ante eventos geopolíticos impredecibles, factores psicológicos y comportamientos irracionales que con frecuencia influyen en las fluctuaciones del mercado. La Inteligencia Artificial Explicable (XAI) es una herramienta útil al proporcionar transparencia y claridad en los modelos predictivos, y permite a los usuarios entender con mayor facilidad la lógica detrás de las decisiones algorítmicas. Este artículo enfatiza la importancia de integrar XAI en sistemas de Trading Algorítmico Ético (TAE), para promover mayor confianza por parte del usuario final, sobre todo los inversionistas. En el artículo se presenta la relevancia de la capa ética para los modelos de TAE, y se enfatiza en los aspectos como la transparencia y explicabilidad, utilizando la importancia de características para Lasso, Ridge y SHAP. Esto permite identificar las variables más representativas que influyen en las predicciones del modelo de TAE, aporta transparencia y facilita el entendimiento sobre cómo se generan dichas predicciones. Finalmente, se propone que el análisis explicativo del modelo y la experiencia humana fortalece la confiabilidad de las predicciones generadas en el TAE.

Palabras clave: Ética, trading algorítmico, inteligencia artificial, transparencia, explicabilidad.

Ethical Algorithmic Trading (TAE): The Value of the Ethical Layer for Artificial Intelligence System

Resumen Algorithmic trading has significantly transformed financial markets by enabling rapid decisions based on predictive machine learning

models. However, these systems often lack transparency and explainability, making it difficult for users to understand how and why certain automated decisions are made. This challenge is exacerbated by unpredictable geopolitical events, psychological factors, and irrational behaviors that frequently influence market fluctuations. Explainable Artificial Intelligence (XAI) is a useful tool by providing transparency and clarity in predictive models, allowing users to more easily understand the logic behind algorithmic decisions. This article emphasizes the importance of integrating XAI into Ethical Algorithmic Trading (EAT) systems to foster greater trust on the part of end-users, especially investors. This article presents the relevance of the ethical layer for TAE models, emphasizing its application in aspects such as transparency and explainability. It uses feature importance for Lasso, Ridge, and SHAP, which allows for the identification of the most representative variables that influence the TAE model's predictions. This provides transparency and facilitates understanding of how these predictions are generated. Finally, it proposes that explanatory analysis of the model and human experience strengthens the reliability of the predictions generated in TAE.

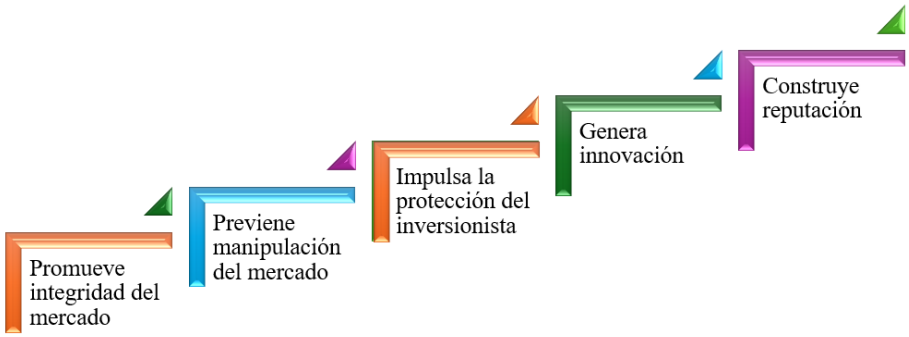
Keywords: Ethics, algorithmic trading, artificial intelligence, transparency, explainability.

1. Introducción

El trading algorítmico, entendido como la aplicación de algoritmos de aprendizaje automático para tomar decisiones financieras rápidas y basadas en datos, ha ganado gran relevancia en los mercados financieros modernos. Sin embargo, uno de los desafíos más importantes de estos sistemas automatizados es la falta de transparencia y explicabilidad sobre cómo llegan a sus predicciones y decisiones. Esta falta de claridad puede ser problemática, especialmente considerando la naturaleza volátil e impredecible de eventos económicos y geopolíticos que afectan significativamente a los mercados financieros.

En este contexto, la denominada inteligencia artificial explicable (XAI) adquiere relevancia, ya que proporciona mecanismos que permiten interpretar con mayor claridad cómo y por qué un modelo de aprendizaje de máquina genera una predicción específica. El objetivo de este artículo es presentar un enfoque centrado en la explicabilidad de modelos predictivos para trading algorítmico, utilizando Lasso, Ridge y SHAP. Estas herramientas permiten no solo seleccionar automáticamente las variables más influyentes en las predicciones del modelo, sino también entender a mayor detalle su impacto individual en la predicción del precio y volumen de las acciones de empresas seleccionadas y listadas en el Nasdaq.

La integración de técnicas de XAI, incluyendo Lasso, Ridge y SHAP en el trading algorítmico ético (TAE), puede contribuir a una mayor transparencia hacia el usuario final o inversionista, facilitando que incluso que usuarios sin conocimientos especializados comprendan cómo se generan las decisiones algorítmicas. Em suma, lo mejor parece ser un enfoque híbrido, que combine



font=footnotesize, labelfont=bf

Fig. 1. Beneficios de incorporar una capa ética en el trading algorítmico. Elaboración con base en Faster Capital [4].

precisión algorítmica con interpretabilidad humana, para promover decisiones de inversión más informadas, responsables y robustas, adaptadas a la realidad compleja y dinámica de los mercados financieros.

2. Trading algorítmico ético

Definiremos en el presente artículo çapa ética"del trading algorítmico como un subproceso de la compra venta de acciones que incorpora un conjunto de consideraciones y principios éticos que guían el diseño, la implementación y el uso de algoritmos en los mercados financieros, tales como el control humano, transparencia, explicabilidad, responsabilidad y fiabilidad. La capa ética en este sentido no solamente aporta un cumplimiento legal o regulatorio, o simplemente un subproceso deseable para los usuarios del trading algorítmico, sino que aporta valor agregado y promueve el desempeño óptimno de los sistemas automatizados de trading al combinarse con la experiencia humana.

Al respecto, Faster Capital reitera que las prácticas comerciales algorítmicas éticas son esenciales para promover la integridad del mercado, proteger a los inversores y garantizar la salud y la estabilidad a largo plazo de los mercados financieros. Al adoptar estas prácticas , los participantes del mercado pueden ayudar a generar confianza en los mercados, lo cual es esencial para su éxito. En la Figura 1 se muestran los beneficios de incorporar una capa ética en el trading algorítmico [4].

2.1. Necesidad de capa ética en el proceso de trading algorítmico

Parte de la necesidad de la capa ética es que la automatización de los mercados financieros que emplea modelos automatizados sigue creciendo, tanto para uso

personal de los usuarios como para uso de plataformas masivas especializadas en trading algorítmico, esta situación genera desafíos importantes en términos de transparencia, seguridad y fiabilidad. Sin una capa ética sólida en el proceso de trading, los algoritmos no están exentos de riesgos, ya que los datos no necesariamente alcanzan a capturar el contexto de variables que afectan el precio de las acciones, ya sea por fenómenos políticos, conflictos bélicos, decisiones irracionales o incluso la manipulación de los mercados financieros.

Al respecto, Gawde señala que por propia naturaleza de çaja negra"de los algoritmos plantea un reto ético en el trading algorítmico, ya que puede ralentizar la rendición de cuentas y conducir a una toma de decisiones con resultados no deseables.

Ciertamente se ha desarrollado una metodología de explicabilidad para enfrentar este problema (XAI), que aborda esta situación brindando más información sobre los procesos de toma de decisiones de los algoritmos, mejorando así la transparencia y la rendición de cuentas [5].

Por otro lado, Gawde advierte que el trading algorítmico en sus inicios solía beneficiar a los grandes inversionistas, lo que amplíaba la brecha entre éstos y los inversionistas individuales más pequeños. Los inversionistas individuales solían experimentar un .efecto de desplazamiento.en entornos donde las operaciones se ejecutaban mediante algoritmos a velocidades y volúmenes inalcanzables para los demás. Esta situación provocó debates relacionados con la equidad del trading algorítmico [5].

De hecho, los algoritmos de alguna forma heredan los desafíos éticos asociados con el diseño y disponibilidad de nuevas tecnologías y aquellos asociados a la manipulación de grandes volúmenes de datos, esto implica que el daño causado por la actividad algorítmica es difícil de depurar, pero también es difícil identifcar quién debería ser considerado responsable por el daño causado [9].

Parra y Cruz señalan que muy pocos traders logran obtener ganancias, dos por ciento de manera impredecible, y apenas uno por ciento de manera predecible. Se podría intepretar que se refieren precisamente a traders .amateurs"que no necesariamente disponen de la capacitación especializada, y la experiencia acumulada para operar de forma efectiva en los mercados de valores. No obstante, estas cifras nos dan una idea del impacto que puede tener el uso indiscriminado de trading algorítmico sin una capa ética robusta que brinde condiciones mínimas de transparencia y explicabilidad de los algoritmos empleados, responsabilidad de los posibles daños cuasados por las pérdidas de las operaciones financieras. En la Figura 2 se muestran los rendimientos de los traders particulares [11].

2.2. Principales principios del trading algorítmico ético

El valor para al usuario deriva no sólo de modelo predictivo del precio y volumen de las acciones de las empresas que cotizan en el Nasdaq, sino de los componentes de la capa ética. Considerando lo señalado por la Unión Europea, se tiene lo siguiente [2]:

1. **Control humano**. Se pueden supervisar, regular e intervenir en los sistemas de inteligencia artificial, y en el caso específico del trading algorítmico, esto

Fig. 2. Rendimientos de Traders Particulares. Elaboración con base en Parra [11].

implica complementar el resultado con intervención humana experta, lo cual puede ayudar a gestionar riesgos inherentes al diseño desarrollo y ejecución de los modelos predictivos usados en el trading. Por un lado, es impoprtante advertir que los algoritmos pueden reaccionar desproporcionadamente a fluctuaciones del mercado, lo que puede desencadenar caídas abruptas de precios y alta volatilidad. Por otro lado, los conjuntos de datos empleados en los modelos, no necesariamente capturan por completo los efectos de los fenómenos geopolíticos, la irracionalidad de los agentes económicos en los mercados financieros, ni la posible manipulación de los mercados financieros a través de anuncios públicos u operaciones de gran volumen que buscan incrementar o reducir el precio de un activo financiero.

2. **Responsabilidad**. Se pueden establecer mecanismos para atribuir responsabilidades legales y éticas ante posibles consecuencias negativas o errores cometidos por los sistemas automatizados del trading algorítmico. En el entorno del trading las decisiones se toman a velocidades rápidas, en ocasiones en milisegundos, por lo que se requiere la asignación de responsabilidades a diseñadores y operadores de los sistemas algorítmicos, para evitar fraudes y errores masivos.

3. **Transparencia y explicabilidad**. Se puede dar claridad sobre cómo funciona un sistema de inteligencia artificial, incluyendo qué datos utiliza, cómo los procesa y cómo toma las decisiones. Ciertamente los modelos de çaja negraçomo las redes neuronales representan un desafío, pero hay modelos Por otro lado, promover que los modelos y algoritmos sean explicables y auditables, puede reforzar la confianza en el trading algorítmico.

3. Implementación del subproceso del trading algorítmico ético

3.1. Recolección y tratamiento de los datos

Para este proyecto en particular, se emplearon los datos de las acciones de Google descargados del sitio de Yahoo Finance. En particular se usaron las variables proporcionadas por Yahoo Finance: precio de apertura, precio de cierre, precio máximo, precio de cierre ajustado, y volumen de transacciones de la acción. No obstante lo anterior, también se emplearon otras variables que se consideró que pudieran tener un impacto en el valor de las acciones: la tasa de interés (que representa el costo de oportunidad de los inversionistas en instrumentos de riesgo vs instrumentos de no riesgo), así como el crecimiento del Producto Interno Bruto.

Para la Comisión para el Mercado Financiero, la tasa de interés se puede definir como el porcentaje del crédito que se paga de manera adicional a la cantidad de dinero (o capital) que se está pidiendo mediante una operación de crédito [1]. Por su parte, la Organización para la Cooperación y el Desarrollo Económicos define el PIB como la medida estándar del valor agregado creado mediante la producción de bienes y servicios en un país durante un periodo determinado [10]. Este indicador también mide los ingresos obtenidos de dicha producción, o la cantidad total gastada en bienes y servicios. Para obtener tanto el PIB como la tasa de interés, se emplearon tanto las librerías de Pyhton como APIs de sitios de Internet.

Al final las variables independientes empleadas en el modelo, para predecir el precio ajustado del activo financiero, son las siguientes:

- **Open:** Precio de apertura del activo financiero en el día.
- **High:** Precio más alto alcanzado por el activo financiero durante el día.
- **Low:** Precio más bajo alcanzado por el activo financiero durante el día.
- **Volume:** Volumen negociado del activo financiero durante el día.
- **GDP:** Producto Interno Bruto (indicador económico general).
- **FEDFUNDS:** Tasa de fondos federales (tasas de interés en EE. UU.).
- **Lag_1, Lag_2, Lag_3, Lag_4, Lag_5:** Valores retrasados (*lags*) del precio del activo financiero, usados para capturar efectos temporales y dependencias del activo en días previos.

3.2. Aspectos relevantes para el ciclo de vida de los datos

La metodología orientada a la capa ética del modelo de aprendizaje de máquina para poder predecir el precio y volumen de las acciones de las empresas que cotizan en el Nasdaq (PRED-NASDAQ), que se basa en los principios éticos prioritarios abarca todo el ciclo de vida de dicho sistema:

- **Generación y almacenamiento de datos de Stock Capital** (Transparencia). Indispensable asegurar transparencia absoluta en los procesos y

subprocesos relacionados con la obtención y almacenamiento de los datos financieros. Para cumplir con este principio, se pueden detallar claramente los cambios, fuentes y métodos utilizados en la adquisición y almacenamiento de los datos del mercado bursátil Nasdaq.

- **Preprocesamiento** (Transparencia). Ciertamente el preprocesamiento de datos implica procedimientos técnicos como limpieza, normalización y selección de variables, los cuales deben ser documentados exhaustivamente para garantizar la transparencia total. Esto incluye el registro y justificación clara de decisiones metodológicas tomadas durante la transformación de los datos originales, asegurando que los usuarios tengan visibilidad sobre cómo se prepara la información antes de alimentar el modelo predictivo.
- **Modelos** (Explicabilidad). En la etapa de construcción y entrenamiento de modelos predictivos, la explicabilidad es muy importante, e incluye explicar claramente cómo y por qué realizan determinadas predicciones, que permitan a los usuarios comprender la lógica interna del modelo, incluyendo el peso y la importancia de cada variable financiera involucrada en las predicciones.
- **Predicciones** (Control humano / Responsabilidad / Rendición de cuentas). Esta etapa reconoce la necesidad de un control humano activo sobre las predicciones generadas por el modelo. Se podría establecer un sistema de validación en el que expertos humanos supervisan y evalúan regularmente los resultados del modelo para identificar anomalías o sesgos potenciales. Además, se puede implementar un mecanismo de rendición de cuentas donde las decisiones predictivas pueden ser auditadas y justificadas para usuarios finales o inversionistas.
- **Interfase con el usuario** (Privacidad). Finalmente, la interfaz que interactúa con el usuario debe asegurar rigurosamente la privacidad de los datos personales y financieros del usuario. Esto implica el cumplimiento estricto de estándares internacionales de protección de datos, implementando políticas robustas de confidencialidad, mecanismos de seguridad avanzados y protocolos claros de gestión del consentimiento del usuario final.

A continuación en la Figura 3, se representa este diagrama de flujo de los procesos del TAE con sus respectivos principios éticos.

4. Un caso de explicabilidad para el trading algorítmico ético

La Unión Europea señala que la inteligencia artificial explicable (XAI) es la capacidad de los sistemas de inteligencia artificial de proporcionar explicaciones claras y comprensibles de sus acciones y decisiones, y su principal objetivo es hacer que el comportamiento de estos sistemas sea comprensible para los seres humanos al dilucidar los mecanismos subyacentes de sus procesos de toma de decisiones [3]. En este caso el objetivo es explicar claramente cómo los modelos llegan a sus predicciones, y al respecto se consideró que Lasso, Ridge y SHAP podían contribuir significativamente a esta explicabilidad.

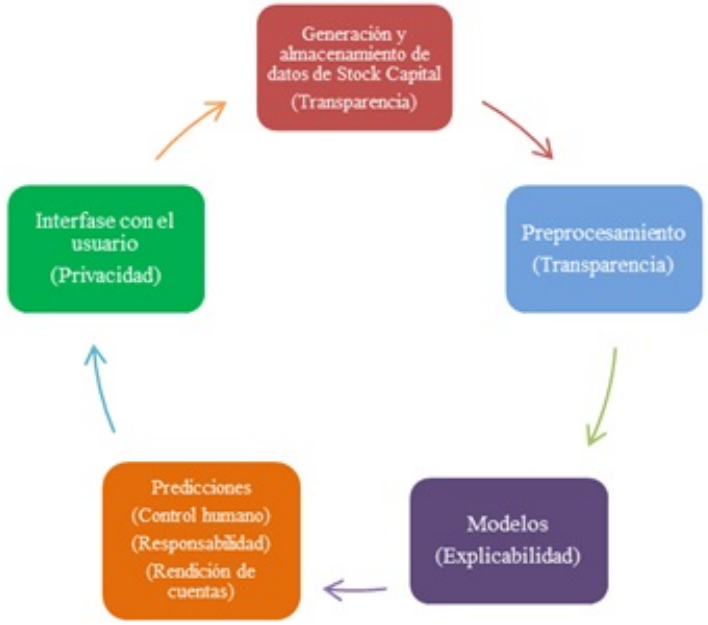


Fig. 3. Diagrama de flujo de los procesos del TAE con sus respectivos principios éticos.

Por su parte, IBM define la regresión Lasso (Least Absolute Shrinkage and Selection Operator) como una técnica de regularización que aplica una penalización para evitar el sobreajuste, mejorar la precisión de los modelos estadísticos, reducir la complejidad y realizar selección automática de características [7]. En ese sentido se puede decir que Lasso facilita la selección automática de las variables más importantes, simplificando el modelo al eliminar variables irrelevantes. A continuación, en la Figura 4 se representa la importancia de características para Lasso.

- **Variables con mayor importancia en Lasso:**
  - **Variable Lag_1.** Tiene la mayor importancia positiva, lo que implica que el precio del activo financiero del día inmediatamente anterior (retraso de 1 día) es altamente relevante para predecir el comportamiento actual o futuro del precio del activo.
  - **Variable Open.** Posee una importancia significativa, destacando que el precio de apertura diario contribuye considerablemente a la predicción del modelo Lasso.
  - **Variable High.** Presenta una importancia considerable, aunque menor que las anteriores, indicando que el precio más alto alcanzado en el día sigue siendo relevante para el modelo.
- **Variables con importancia moderada, baja o nula en Lasso:**

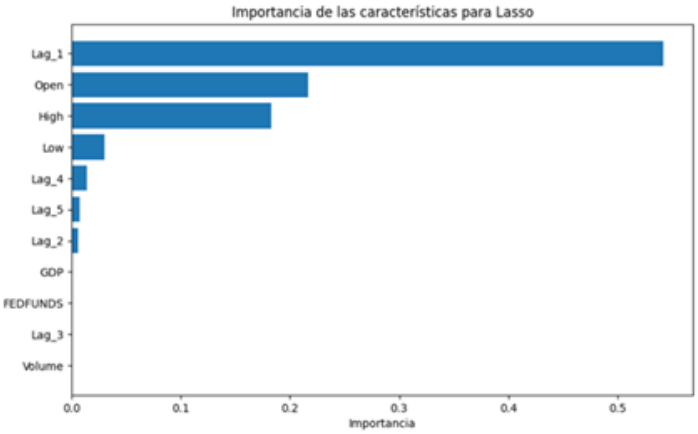


Fig. 4. Importancia de características para Lasso

- **Variable Low.** Aunque tuvo mucha relevancia en Ridge, presenta ahora una importancia moderada en Lasso, pero aún es considerada por el modelo.
- **Variables Lag_4, Lag_5, Lag_2.** Estas variables retrasadas muestran importancias muy pequeñas, aunque no nulas, lo que significa que el modelo aún las toma en cuenta en menor medida.
- **Variables macroeconómicas (GDP, FEDFUNDS) y volumen negociado (Volume).** Totalmente penalizadas (reducidas a cero), lo que significa que no aportan información significativa adicional según el modelo Lasso en este análisis particular.

Asimismo, Murel y Kavlakogl mencionan que la regresión Ridge es una técnica de regularización estadística que corrige el sobreajuste de los datos de entrenamiento en los modelos, esto permite reducir los errores causados por el sobreajuste y puede mejorar la generalización del modelo [8]. A diferencia de Lasso, Ridge no elimina completamente variables, sino que reduce sus coeficientes de forma proporcional. Al respecto, se puede advertir que Ridge puede mejorar la estabilidad del modelo, asignando pesos proporcionales a las variables relevantes sin eliminarlas por completo. A continuación, en la Figura 5 se representa la importancia de características para Ridge.

- **Variables con importancia positiva alta:**
  - **Low (precio mínimo diario):**
    - Tiene la mayor importancia positiva, por lo que un precio mínimo más alto durante el día incrementa la predicción del precio ajustado de cierre.
    - Sugiere que el precio mínimo del día es muy relevante para predecir movimientos futuros positivos del activo.
  - **High (precio máximo diario):**

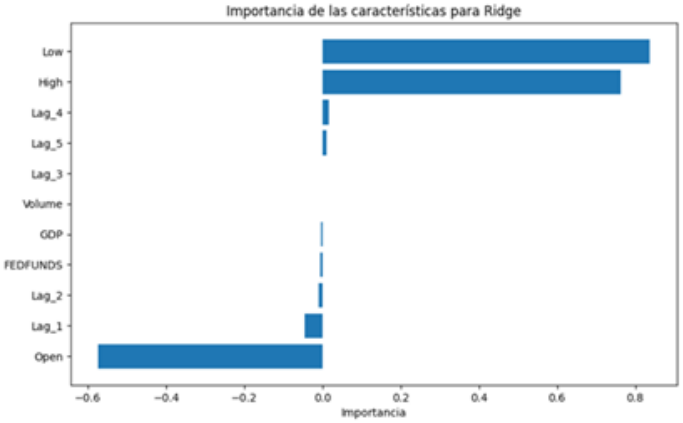


Fig. 5. Importancia de características para Ridge

  - Segunda importancia positiva más alta, lo cual indica que un precio máximo elevado durante la sesión contribuye positivamente a las predicciones del modelo Ridge.
  - Confirma que los precios más altos alcanzados en el día tienen una fuerte relación positiva con el precio ajustado de cierre.
- **Variables con importancia negativa alta:**
  - **Open (precio de apertura diario):**
    - La variable con la importancia negativa más significativa, lo que implica una relación inversa con el precio ajustado de cierre.
    - Sugiere que altos precios de apertura podrían asociarse con caídas posteriores del precio ajustado al cierre del día.
- **Variables con importancia muy baja (casi nula):**
  - **Lag_1, Lag_2, Lag_3, Lag_4, Lag_5:**
    - Estas variables temporales o retrasadas (precio de días previos) son poco significativas para Ridge.
    - Ridge reduce sus coeficientes, indicando que la dependencia temporal inmediata no es tan relevante en presencia de otras variables como *High* y *Low*.
  - **GDP, FEDFUNDS, Volume:**
    - Estas variables económicas y de volumen tienen importancias prácticamente nulas, lo cual significa que en el modelo Ridge se consideran poco relevantes en la predicción de los movimientos del activo financiero.

Finalmente, Guetta señala que SHAP (Shapley Additive Explanations) es un enfoque de explicabilidad del aprendizaje automático para comprender la importancia de las características en instancias individuales, y proporciona una comprensión detallada de las explicaciones locales [6]. Al asignar un valor numérico a la influencia de cada característica, SHAP de alguna manera facilita

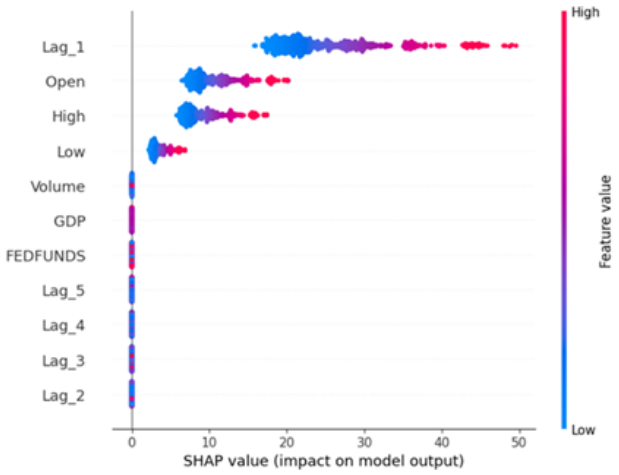


Fig. 6. Gráfica SHAP indicando la importancia relativa de cada variable sobre la predicción del precio ajustado al cierre (*Adj Close*).

la interpretación de las decisiones del modelo, lo que puede promover la confianza y facilitar la identificación de posibles sesgos.

La Figura 6 sobre SHAP, indica cuánto y cómo cada variable afecta las predicciones del precio ajustado al cierre (*Adj Close*) en el modelo predictivo.

Al respecto, tenemos que:

- **Variable Lag_1 (precio del día anterior).** La variable más determinante, donde altos valores del precio anterior contribuyen positivamente a predicciones más altas, y destaca la fuerte dependencia del precio actual respecto al precio inmediato anterior.
- **Variable Open (precio de apertura) y High (precio máximo).** También tienen alta relevancia, y los valores más altos de estas variables generalmente implican mayores predicciones del modelo sobre el precio ajustado de cierre.
- **Variable Low (precio mínimo).** Tiene relevancia moderada, donde los altos valores tienden a contribuir positivamente, pero menos significativamente que las anteriores.
- **Variables con baja relevancia:** variables económicas como *GDP* y *FEDFUNDS*, así como *Volume* y lags adicionales (*Lag_2, Lag_3, Lag_4, Lag_5*), tienen poco impacto sobre la predicción, indicando su baja relevancia para el modelo.

5. Recomendaciones y ventajas de aplicación

Como se pudo observar, la explicabilidad o XAI puede ser muy útil en el desarrollo de modelos predictivos financieros, especialmente en el trading algorítmico, ya que además de que es deseable que los modelos tengan predicciones

precisas, también es esencial que los usuarios o inversionistas comprendan cómo y por qué se generan dichas predicciones. Esto es clave para garantizar la transparencia, la confianza y la toma de decisiones, es por ello que Lasso, Ridge y SHAP, se perfila como una buena opción para identificar y explicar las variables más relevantes que impactan en la predicción del precio de las acciones de las empresas.

En este caso el uso de Lasso permitió hacer una selección automática de características, pudiendo eliminar aquellas variables cuya influencia en el modelo es mínima o nula. Es por ello que no solo ayuda a optimizar el rendimiento del modelo y reducir su complejidad, sino que también facilita la interpretabilidad, ya que permite identificar las variables más significativas para las predicciones del precio ajustado de los activos financieros.

Por otro lado, Ridge permitió identificar aquellas variables que, aunque menos influyentes, aún aportan valor a las predicciones, lo que facilita el análisis y ayuda a identificar cómo diferentes factores afectan las decisiones de inversión. También el uso de SHAP ofreció un nivel de explicabilidad muy específico, ya que permitió visualizar el impacto individual de cada característica, lo que permite reforzar la transparencia del modelo.

Finalmente, cabe señalar que aquí se mencionó solamente el caso de explicabilidad y su aplicación en modelos de trading algorítmico, pero se abren muchas posibilidades para explorar nuevas aplicaciones relacionadas con otros principios éticos relevantes relacionados como el control humano, para promover que las decisiones automatizadas sean complementadas y supervisadas por expertos humanos en el mercado financiero. También la posibilidad de crear subprocesos asociados a la responsabilidad y rendición de cuentas para definir claramente quién es responsable respecto a la toma de decisiones, así como poder justificar adecuadamente esa toma de decisiones. La integración de estos principios y otros más, permitirá proteger a los inversionistas y fortalecer la confianza en el trading algorítmico.

Referencias

1. CMF. ¿En qué consiste la Tasa de Interés? Disponible en: https://www.cmfchile.cl/educa/621/w3-article-27169.html#:~:text=La%20tasa%20de%20inter%C3%A9s%20es%20el%20resultado%20de%20un%20c%C3%A1lculo,mediante%20una%20operaci%C3%B3n%20de%20cr%C3%A9dito, último acceso: 2024/02/19.
2. European Commission. Ethics Guidelines for Trustworthy AI. Disponible en: https://digital-strategy.ec.europa.eu/es/library/ethics-guidelines-trustworthy-ai, último acceso: 2024/02/19.
3. European Data Protection Supervisor (EDPS). TechDispatch on Explainable Artificial Intelligence (XAI) (2023). Disponible en: https://www.edps.europa.eu/system/files/2023-11/23-11-16_techdispatch_xai_en.pdf, último acceso: 2024/03/21.
4. Faster Capital. The Importance Of Ethical Algorithmic Trading Practices. Disponible en: https://fastercapital.com/topics/

the-importance-of-ethical-algorithmic-trading-practices.html, último acceso: 2024/03/21.
5. Gawde, A., Jawale, A. Ethical Considerations in Algorithmic Trading: Recent Developments, Challenges, and the Path Forward. International Journal of Creative Research Thoughts 12(12), a91–a105 (2024)
6. Guetta, N. SHAP Global Explanations for Machine Learning. Disponible en: https://www-aporia-com.translate.goog/learn/explainability/shap-global-explantations-ml/?_x_tr_sl=en&_x_tr_tl=es&_x_tr_hl=es&_x_tr_pto=wa, último acceso: 2024/02/09.
7. IBM. Lasso Regression (2024). Disponible en: https://www.ibm.com/mx-es/think/topics/lasso-regression, último acceso: 2024/03/21.
8. Murel, J., Kavlakoglu, E. Ridge Regression (2023). Disponible en: https://www.ibm.com/es-es/think/topics/ridge-regression, último acceso: 2024/03/21.
9. Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., Floridi, L. The Ethics of Algorithms: Mapping the Debate. Big Data & Society 3(2) (2016)
10. OCDE. El Producto Interno Bruto o Producto Interior Bruto (PIB) (2024). Disponible en: https://www.oecd.org/espanol/estadisticas/pib-espanol.htm#:~:text=El%20Producto%20Interno%20Bruto%20o%20Producto%20Interior%20Bruto%20(PIB)%20es,pa%C3%ADs%20durante%20un%20periodo%20determinado, último acceso: 2024/03/15.
11. Parra, M. Estadísticas de Trading. NewTrading (2025). Disponible en: https://www.mynewtrading.com/convertirse-en-trader-rentable/, último acceso: 2025/03/09.

ISSN 1870-4069

Composición asistida por red neuronal para piano en jazz

Ismael Medina Muñoz

Microsoft, Redmond,
USA

ismedina@microsoft.com

Resumen. La inteligencia artificial ha asumido un papel importante en actividades que anteriormente se consideraban exclusivamente humanas. La IA generativa es un área de investigación vibrante, con un creciente interés en campos de aplicación relacionados con las artes. Esta investigación tiene como objetivo demostrar una red neuronal recurrente para crear nuevas secuencias de n-notas a partir de un conjunto inicial de n-notas, y se aplica un enfoque probabilístico para establecer la duración de cada nota en el conjunto resultante. Esto sirve como asistente en la composición de piano para melodías de jazz, produciendo una representación simbólica de la música en forma de partituras. Al aprovechar las técnicas de aprendizaje automático, la IA puede analizar patrones y estructuras en composiciones de jazz existentes, permitiéndole generar secuencias nuevas y armoniosas. Las secuencias generadas pueden ser luego refinadas y ajustadas por compositores humanos, resultando en un proceso colaborativo entre la IA y la creatividad humana. Esta integración de la IA en el proceso creativo abre nuevas posibilidades para la innovación y exploración musical.

Palabras clave: Recurrent neural network, composición musical, asistente.

Neural Network-Assisted Composition for Piano in Jazz

Abstract. Artificial intelligence has taken on an important role in activities that were previously considered exclusively human. Generative AI is a vibrant area of research, with growing interest in application fields related to the arts. This research aims to demonstrate a recurrent neural network for creating new sequences of n-notes from an initial set of n-notes, and a probabilistic approach is applied to establish the duration of each note in the resulting set. This serves as an assistant in piano composition for jazz melodies, producing a symbolic representation of the music in the form of scores. By leveraging machine learning techniques, AI can analyze patterns and structures in existing jazz compositions, allowing it to generate novel and harmonious sequences. The generated sequences can then be further refined and adjusted by human composers,

resulting in a collaborative process between AI and human creativity. This integration of AI into the creative process opens up new possibilities for musical innovation and exploration.

Keywords: Recurrent neural network, musical composition, assistant.

1. Los datos de entrenamiento de la red neuronal

1.1. Partituras como representación simbólica de la música

Un gran número de partituras producidas con el software libre **MuseScore** están disponibles de forma gratuita. Se descargaron partituras de jazz ejecutadas en piano que posteriormente fueron analizadas programáticamente utilizando la librería **MS3**, desarrollada para Python por Johannes Hentschel [2].

Los datos que las partituras proporcionaron reflejan que un ejecutante de piano utiliza ambas manos para ejecutar piezas de jazz. Esto se representa mediante 2 pentagramas, el superior sirve para describir las notas pulsadas en la clave de Sol por la mano derecha, mientras que el inferior sirve para describir las notas pulsadas en la clave de Fa con la mano izquierda. Se demostró que la mano izquierda tiende a tocar teclas con sonidos más graves mientras que la mano derecha sirve para tocar teclas con sonidos más agudos, con tendencia a pulsar un máximo de 5 teclas al mismo tiempo por cada mano. Para aquellas excepciones, donde cada mano pulsó más de 5 teclas se tomaron sólo las primeras 5 notas de forma ascendente de acuerdo con su valor del Musical Instrument Digital Interface (MIDI), donde los valores menores representan sonidos más graves mientras que los valores mayores representan sonidos más agudos.

A las notas pulsadas al mismo tiempo para cada mano se le denominó como "rebanada" de n-notas. Así pues, la mano izquierda registra que en el pentagrama en clave de Fa (staff 2) pulsa 5 notas a lo más al mismo tiempo, mientras que la mano derecha registra que en el pentagrama en clave de Sol (staff 1) pulsa 5 notas a lo más al mismo tiempo (vea la figura 1). Una rebanada de 0-notas en ambas manos representa el silencio en la música.

Las notas pulsadas están descritas por su valor MIDI (vea la figura 2). Estas rebanadas de n-notas de la mano derecha e izquierda pulsadas al mismo tiempo se codificaron para crear el vocabulario, como si de palabras de algún lenguaje se trataran y su representación se da por un identificador numérico.

1.2. Secuencias de rebanadas"de n-notas como palabras de una oración

Con las rebanadas de n-notas ya codificadas con un identificador fue ya posible codificar también secuencias de rebanadas de n-notas como si se tratasen de palabras en una oración. Para esta investigación se generaron secuencias de 3 palabras (vea la figura 3).

Cabe aclarar que se removieron rebanadas subsecuentes que representan a la misma palabra para evitar que la red neuronal aprenda a producir secuencias repetitivas de las mismas notas.

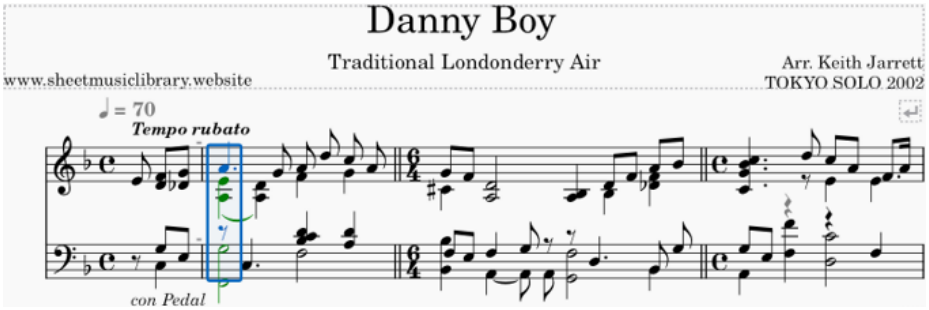


Fig. 1. Rebanada"de 3-notas para la mano derecha y de 2-notas para la mano izquierda.

staff01_note01	staff01_note02	staff01_note03	staff01_note04	staff01_note05	staff02_note01	staff02_note02	staff02_note03	staff02_note04	staff02_note05
0	0	0	0	0	0	0	0	0	0
71	0	0	0	0	0	0	0	0	0
71	0	0	0	0	36	48	0	0	0
0	0	0	0	0	31	43	0	0	0
0	0	0	0	0	36	48	0	0	0

Fig. 2. Rebanada"de n-notas para cada mano (staff 1 o staff 2) y representación de las teclas pulsadas con su valor MIDI

	prev_n_notes_id	actual_n_notes_id	next_n_notes_id
0	0	1	0
1	0	2	3
2	2	3	4
3	3	4	3
4	4	3	4
...	...	...	...
78846	179	18226	0
78847	18226	179	311
78848	311	178	18228
78849	178	18228	0
78850	18228	0	0

Fig. 3. Secuencias únicas de 3 rebanadas"de n-notas.

1.3. La duración de las notas y la función de masa de probabilidades

El proceso de construcción del vocabulario también produjo información fuertemente asociada a la nota pulsada en cada pentagrama, la duración de dichas notas. La duración de cada nota se contabilizó para producir distribuciones de probabilidad de las notas de acuerdo con la cantidad de n-notas pulsadas por

Tabla 1. Diferencia de distribución de rebanadas de n-notas.

Métrica	Staff 1	Staff 2
2/2	p-value = 0.000, K-S 2c = 0.4269	p-value = 0.000, K-S 2c = 0.8958
2/4	p-value = 0.000, K-S 2c = 0.4164	p-value = 0.000, K-S 2c = 0. 8958
4/4	p-value = 0.000, K-S 2c = 0.2564	p-value = 0.000, K-S 2c = 0. 8958
3/4	p-value = 0.000, K-S 2c = 0.3010	p-value = 0.000, K-S 2c = 0. 8958
5/4	p-value = 0.000, K-S 2c = 0.4988	p-value = 0.000, K-S 2c = 0. 8958

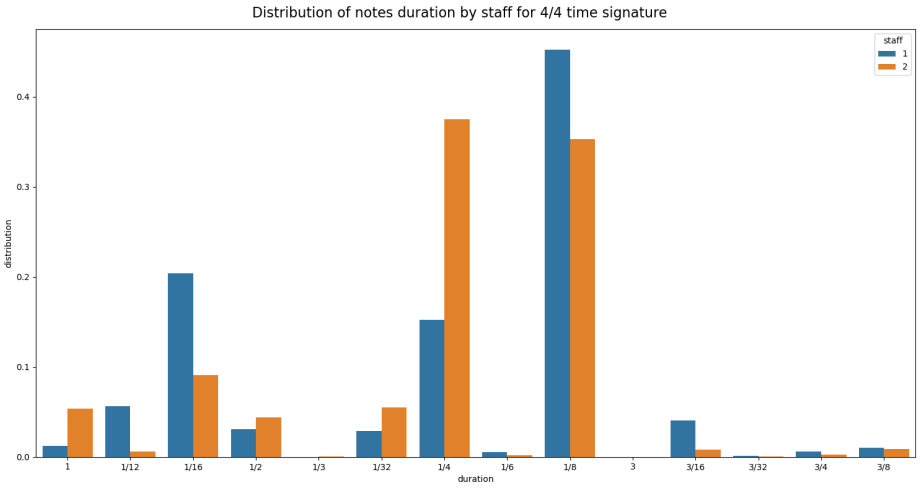


Fig. 4. Distribución de la duración de las notas por cada mano (staff 1 y 2) para una pieza en métrica de 4/4

cada mano, agrupadas por la métrica del pentagrama. Estos datos produjeron funciones de masa de probabilidad de donde se toman valores aleatorios que complementan a las notas producidas por la red neuronal (vea la figura 4).

2. El modelo resultante

Las secuencias de rebanadas"de n-notas como palabras de una oración fueron codificadas como tensores de entrada. Este modelado de secuencias y pilas para entrenar una red neuronal recurrente que produce secuencias fue tomado de las ideas de [1] , como es similar a secuencias de palabras en el procesamiento de lenguaje natural (vea la figura 5).

Se entrenaron dos redes neuronales, una basada en arquitectura LSTM y otra basada en arquitectura GRU. La red con arquitectura GRU se desempeñó mejor durante el entrenamiento y esta fue utilizada para producir nuevas secuencias de n-notas (vea la figura 6).

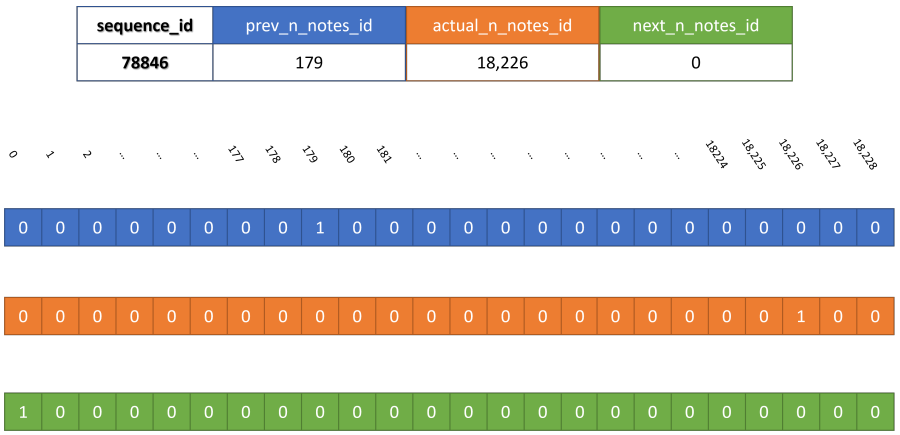


Fig. 5. Descripción del tensor de entrada que presenta un formato de secuencias de palabras

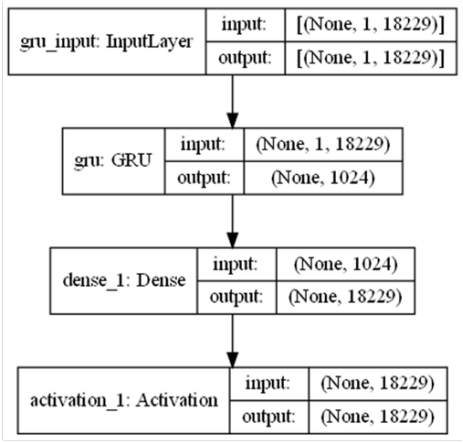


Fig. 6. La arquitectura de la red neuronal entrenada

3. Resultados y conclusiones

Con el modelo elegido, se generaron 100 nuevas secuencias de 64 rebanadas"de n-notas con duraciones de nota tomadas de forma aleatoria de las distribuciones de masa de probabilidad construidas en el experimento. Dichas secuencias fueron representadas en cinco métricas de partitura que luego fueron analizadas y comparadas contra el conjunto de partituras usadas para el entrenamiento utilizando la prueba de Kolmogórov-Smirnov de 2 colas (estadístico K-S 2c).

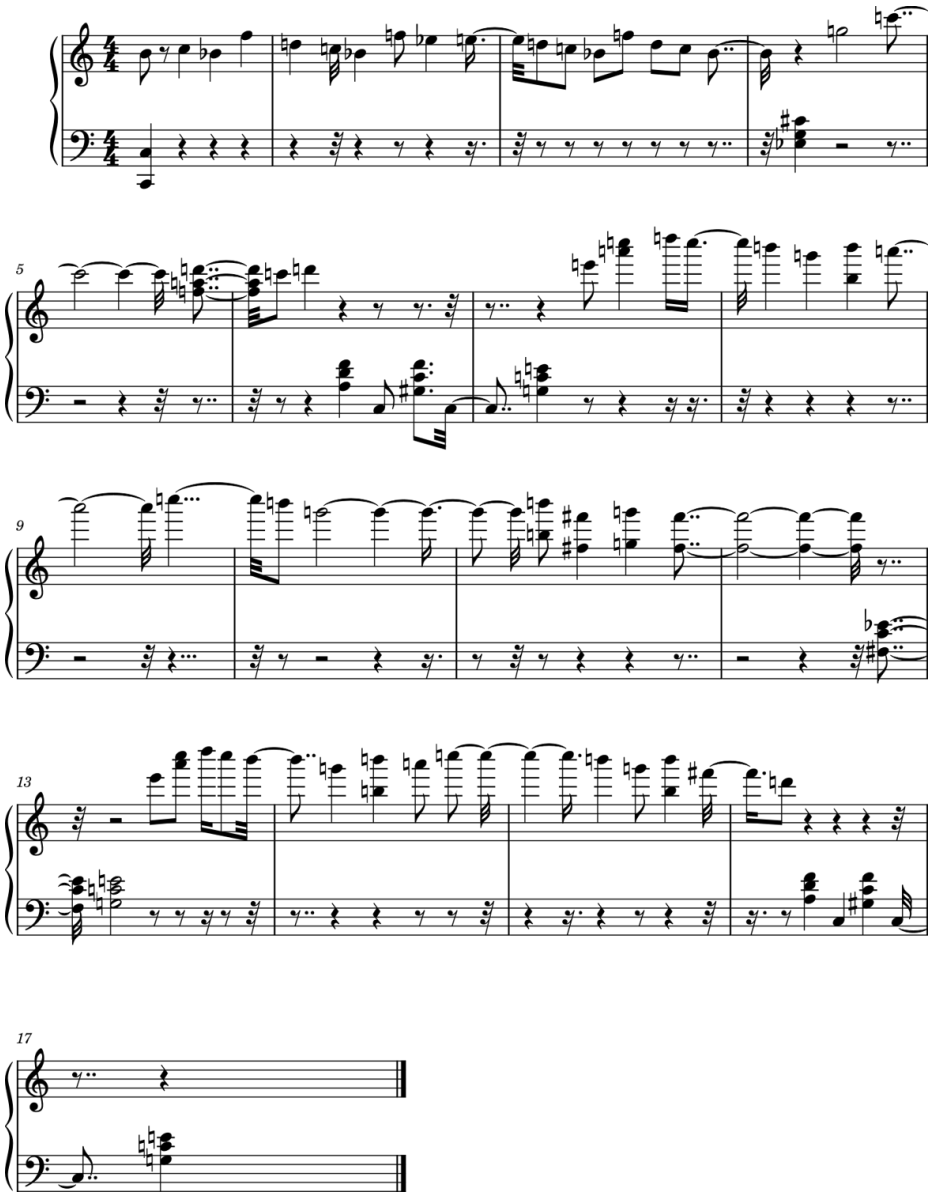


Fig. 7. Archivo PNG creado utilizando la respuesta JSON de la REST API del modelo

Los resultados muestran que las nuevas piezas son estadísticamente diferentes respecto de las partituras de entrenamiento (vea la tabla 1).

El modelo fue desplegado como una Representational State Transfer Application Programming Interface (REST API) que produce secuencias de n-notas con

sus duraciones en formato Comma-separated values (CSV), Extensible Markup Language (XML), MIDI y Portable Network Graphics (PNG). La representación PNG es útil para los ejecutantes de piano (vea la figura 7).

Esto constituye a la red neuronal definida como un asistente para la composición musical de piezas de jazz que brinda ideas nuevas para la composición.

Referencias

1. Goldberg, Y.: Neural Network Methods for Natural Language Processing. pp. 164–166. Springer International Publishing, Cham (2017)
2. Manual, https://johentsch.github.io/ms3/build/html/manual/index.html, last accessed 2023/02/17

ISSN 1870-4069

Explainable Artificial Intelligence (XAI) for Hate Speech Detection Using Social Media Discourse

Muhammad Ahmad, Ildar Batyrshin, Grigori Sidorov*

Instituto Politécnico Nacional (IPN),
Centro de Investigación en Computación (CIC), Mexico City,
Mexico

sidorov@cic.ipn.mx

Abstract. The proliferation of hate speech on social media platforms has escalated into a critical threat to online safety and societal harmony. Traditional deep learning models used for hate speech detection often operate as black boxes, offering no insight into their decision-making processes. This lack of transparency undermines user trust and hinders real-world deployment, especially in sensitive applications such as content moderation. To address this limitation, this study proposes an explainable deep learning model for hate speech detection in Urdu, a low-resource and morphologically rich language. The proposed framework leverages the UA-HSD-2025 dataset of Urdu tweets, applying tailored preprocessing and data augmentation to improve data quality. A Bidirectional Long Short-Term Memory (BiLSTM) network integrated with FastText embeddings captures subword information and contextual dependencies effectively. Experimental results show that the proposed model achieves 99.3% accuracy, significantly outperforming baseline approaches. To enhance interpretability, Local Interpretable Model-agnostic Explanations (LIME) are incorporated to provide word-level explanations for each prediction, transforming the black-box model into a transparent one. In conclusion, this study presents a highly accurate and explainable deep learning solution for Urdu hate speech detection, promoting accountability and trust in automated moderation systems for low-resource languages.

Keywords: Hate speech detection, explainable AI, social media, NLP, deep learning.

1 Introduction

Hate speech detection has emerged as a critical research area within natural language processing (NLP) due to the rapid growth of user-generated content on social media platforms [12,19,18]. Online spaces such as Twitter, Facebook, and Reddit have become central to communication, but they also facilitate the spread of abusive, offensive, and harmful language targeting individuals or groups

based on attributes such as race, religion, gender, or ethnicity [6,13,8]. The proliferation of such content can lead to serious societal consequences, including the reinforcement of stereotypes, psychological harm, and the escalation of real-world conflicts.

Traditional approaches to detecting hate speech relied heavily on manual moderation and rule-based systems, which are often insufficient due to the scale, diversity, and evolving nature of online language [4,11,16]. To address these challenges, researchers have increasingly turned to machine learning and deep learning techniques. Early methods utilized classical algorithms such as Support Vector Machines (SVM) and Naïve Bayes with handcrafted features like bag-of-words and TF-IDF representations [23,1,5]. However, these approaches often struggle to capture contextual nuances, sarcasm, and implicit forms of hate speech.

Recent advancements in deep learning, particularly transformer-based models such as BERT, RoBERTa, and GPT, have significantly improved the performance of hate speech detection systems [17,7,15]. These models leverage contextual embeddings and large-scale pretraining to better understand linguistic subtleties and semantic relationships in text. Furthermore, the integration of explainable artificial intelligence (XAI) techniques has become increasingly important to ensure transparency and trust in automated decision-making systems.

Despite significant advancements, hate speech detection remains a challenging task due to issues such as data imbalance, domain dependency, multilingual variations, and the ambiguity between hate speech and offensive language. Moreover, many state-of-the-art deep learning models, including BERT, operate as black-box systems, limiting their interpretability and reducing user trust in automated decisions. To address these challenges, this study proposes an explainable hybrid deep learning framework for hate speech detection. The proposed approach combines the strengths of deep learning models with pre-trained word embeddings to enhance classification performance. Furthermore, Explainable Artificial Intelligence (XAI) techniques are integrated to provide transparency in model predictions and improve interpretability. This framework aims to achieve robust, accurate, and interpretable hate speech detection, thereby contributing to safer, more reliable, and trustworthy online environments.

2 Related Work

The rapid expansion of social media platforms has intensified the spread of hate speech, raising critical concerns about digital safety, fairness, and ethical content moderation. This has led to a growing body of research focusing not only on improving detection performance but also on enhancing transparency, robustness, and multilingual capability of hate speech detection systems.

Ribeiro et al. [9] investigated transparency and accountability in hate speech detection, particularly in misogyny-related studies, by analyzing annotator documentation across 25 research papers. They proposed a structured framework consisting of six key dimensions, including annotator demographics, training,

and expertise, and introduced a weighted Annotator Metadata Transparency (AMT) score. Their findings revealed significant transparency gaps, especially in demographic and expertise reporting, and an interesting inverse relationship between model performance and transparency, where higher F1 scores often corresponded to poorer documentation quality.

Building on the need for interpretability, Eilertsen et al. [10] addressed the black-box nature of deep learning models by introducing Supervised Rational Attention (SRA). This framework aligns model attention with human-annotated rationales, enabling the model to focus on meaningful linguistic cues during classification. Evaluations on benchmark datasets demonstrated that SRA improves explainability while maintaining competitive performance and fairness, highlighting the importance of integrating human reasoning into model design.

In contrast, Xu et al. [22] focused on the challenges of detecting hate speech in Chinese social networks, where users often employ cloaking techniques to evade detection. They proposed MMBERT, a multimodal framework based on BERT, integrating text, speech, and visual modalities through a Mixture-of-Experts architecture. With a progressive training strategy and modality-specific routing, MMBERT achieved superior performance over both fine-tuned BERT and large language models, demonstrating robustness against adversarial and multimodal inputs.

Similarly, Mishra et al. [14] conducted a comparative study of traditional and advanced models for hate speech detection. They evaluated CNNs, LSTMs, and transformer-based models such as BERT, alongside hybrid architectures. Additionally, they explored text transformation techniques aimed at converting offensive content into neutral expressions. Their findings suggest that while transformer models achieve higher accuracy, hybrid and transformation-based approaches offer valuable improvements in real-world mitigation scenarios.

From a data-centric perspective, Umansky et al. [20] examined the role of annotation quality in fine-tuning large language models, specifically GPT-4o-mini. Using datasets labeled by annotators with varying expertise, they found that only high-quality annotations—particularly from trained experts—improve model performance. Low-quality annotations, in contrast, may degrade detection effectiveness, emphasizing that data quality is more critical than model complexity alone.

Focusing on ensemble learning, Aksoy et al. [3] addressed hate speech detection against LGBTQ+ individuals in Turkish tweets. They fine-tuned multiple large language models and proposed "Chosen Deep," an ensemble approach combining soft and hard voting strategies. Evaluated against traditional models and GPT-4, their method consistently outperformed baselines, demonstrating the effectiveness of ensemble learning for improving classification accuracy.

Addressing multilingual challenges, Ahmad et al. [2] focused on Arabic and Urdu hate speech detection by introducing the UA-HSD-2025 dataset, enriched with binary and multi-class annotations. They explored both translation-based and joint multilingual strategies, evaluating them using traditional machine learning, deep learning, and transformer models such as XLM-R. Their results

confirmed that multilingual transformer models significantly outperform conventional approaches across both languages.

Extending this direction, Usman et al. [21] developed a trilingual hate speech framework covering English, Spanish, and Urdu. They introduced a manually annotated dataset and conducted extensive experimentation using 41 different model configurations, including machine learning, deep learning, and transformer-based methods. Their study demonstrated that GPT-3.5 achieved the best performance, outperforming strong baselines such as XLM-R, particularly in low-resource languages like Urdu.

3 Study Design

3.1 Dataset Collection

The effectiveness of any hate speech detection system largely depends on the quality and representativeness of the underlying dataset. In this study, we utilize the UA-HSD-2025 dataset introduced by Ahmad et al. [2], which was originally developed for multilingual hate speech detection in Arabic and Urdu social media content. The dataset consists of manually annotated tweets collected from Twitter, ensuring real-world linguistic variability and contextual richness.

From the original dataset, we specifically focus on the Urdu subset to address the challenges of low-resource language processing. This subset contains 1,518 samples, where each instance is labeled into two classes: hate and not hate. This binary classification setup allows for a clear and structured formulation of the hate speech detection task.

The selected Urdu dataset is particularly challenging due to the presence of informal language, cultural expressions, and contextual ambiguity commonly observed in social media text. By leveraging this dataset, our study aims to evaluate the performance of advanced machine learning and deep learning models in a realistic low-resource scenario, contributing to more effective and inclusive hate speech detection systems.

3.2 Data Preprocessing

Data preprocessing plays a crucial role in improving the quality and consistency of textual data before model training. In this study, we applied a comprehensive preprocessing pipeline to the Urdu hate speech dataset. Initially, all text samples were converted to lowercase to ensure uniformity and reduce redundancy caused by case sensitivity. Subsequently, we removed noise such as URLs, user mentions, hashtags, punctuation marks, numbers, and special characters commonly present in social media content from Twitter.

Furthermore, stopwords were eliminated to reduce irrelevant linguistic information, and extra whitespace was normalized. Given the informal nature of social media language, additional cleaning steps were applied to handle elongated words, repeated characters, and informal spellings. These preprocessing steps ensured that the dataset was refined and suitable for effective feature extraction and model training.

3.3 Data Augmentation

To address data sparsity and improve model generalization, data augmentation techniques were employed on the training set. Since the dataset is relatively small and imbalanced, augmentation helps in generating diverse linguistic patterns while preserving semantic meaning.

We applied text-based augmentation strategies such as synonym replacement and paraphrasing to create additional training samples for the minority class. This approach helps the model better learn contextual variations of hate speech expressions. In addition, random perturbation techniques, including word shuffling within limited boundaries, were used to increase robustness against lexical variations.

By incorporating these augmentation methods, the dataset was enriched, leading to improved model stability and better generalization performance, particularly in handling unseen or noisy social media text.

4 Application of Models

In this study, we employ a combination of classical machine learning and deep learning models to effectively detect hate speech in Urdu social media text. Specifically, we utilize three machine learning algorithms—Decision Tree (DT), Random Forest (RF), and Support Vector Machine (SVM)—alongside two deep learning architectures, namely Convolutional Neural Network (CNN) and Bidirectional Long Short-Term Memory (BiLSTM).

The choice of feature representation differs based on the nature of the models. For machine learning models, textual data is transformed using Term Frequency–Inverse Document Frequency (TF-IDF). TF-IDF is selected due to its ability to convert text into sparse, fixed-length feature vectors while effectively capturing the importance of words based on their frequency in a document and rarity across the corpus. This makes it highly suitable for traditional classifiers such as DT, RF, and SVM, which perform well on structured and high-dimensional sparse data.

For deep learning models, we use dense word representations that preserve semantic and contextual relationships between words. Therefore, word embedding techniques such as GloVe and FastText are employed. These embeddings map words into continuous vector spaces where semantically similar words are positioned closer together. Unlike TF-IDF, word embeddings capture contextual meaning, syntactic relations, and semantic similarity, making them more effective for sequential models like CNN and BiLSTM.

The mathematical formulation of TF-IDF is defined as follows:

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}}, \tag{1}$$

where $f_{t,d}$ represents the frequency of term t in document d.

$$IDF(t) = \log\left(\frac{N}{1+n_t}\right), \tag{2}$$

Where N is the total number of documents and n_t is the number of documents containing term t.

The final TF-IDF representation is given by:

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t). \tag{3}$$

For word embeddings, GloVe learns word representations by factorizing the global word co-occurrence matrix. Its objective function is defined as:

$$J = \sum_{i,j=1}^{V} f(X_{ij}) \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij}\right)^2, \tag{4}$$

where X_{ij} represents the co-occurrence matrix and w_i, $\tilde{w}_j$ are word vectors.

FastText further enhances word representations by incorporating subword information:

$$w_g = \frac{1}{|G|} \sum_{g \in G} z_g, \tag{5}$$

where g denotes character n-grams and z_g represents subword embeddings.

The Convolutional Neural Network (CNN) is used to extract local feature patterns from embedded sequences. The convolution operation is defined as:

$$c_i = f(W \cdot x_{i:i+k-1} + b), . \tag{6}$$

where $x_{i:i+k-1}$ is the input window, W is the filter matrix, and f is an activation function.

To capture long-range dependencies, we employ a Bidirectional Long Short-Term Memory (BiLSTM) network. The forward and backward hidden states are computed as:

$$\overrightarrow{h_t} = LSTM(x_t, \overrightarrow{h_{t-1}}), \tag{7}$$

$$\overleftarrow{h_t} = LSTM(x_t, \overleftarrow{h_{t+1}}). \tag{8}$$

The final representation is obtained by concatenation:

$$h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]. \tag{9}$$

Overall, the combination of TF-IDF-based machine learning models and embedding-based deep learning models ensures a comprehensive evaluation of both shallow and deep representations for effective hate speech detection.

4.1 Explainability Using LIME

To enhance the interpretability of the proposed hate speech detection framework, we employ Local Interpretable Model-agnostic Explanations (LIME) for model explanation. Since deep learning models such as CNN and BiLSTM operate as black-box systems, their decision-making process is often not transparent. LIME addresses this limitation by providing local explanations for individual predictions, making the model behavior more interpretable and trustworthy.

LIME works by approximating the complex model locally with an interpretable surrogate model. Given an input text instance, LIME perturbs the input data by generating multiple slightly modified samples and observes the corresponding predictions of the trained model. It then assigns weights to these samples based on their proximity to the original instance and fits a simple linear model to approximate the local decision boundary.

In this study, LIME is applied to identify the most influential words contributing to the classification of a text instance as hate or non-hate. This allows us to highlight key linguistic features that drive model predictions, thereby improving transparency and user trust in the system. By visualizing word-level contributions, LIME helps bridge the gap between high-performance deep learning models and their interpretability requirements in sensitive tasks such as hate speech detection.

5 Overall Strategy

This study follows a structured and systematic pipeline for hate speech detection in Urdu social media text. The overall workflow begins with dataset selection from the UA-HSD-2025 corpus, followed by rigorous preprocessing steps including text cleaning, normalization, and noise removal to ensure data consistency. To enhance model generalization, data augmentation techniques are applied to increase data diversity and address class imbalance.

After preprocessing, the dataset is transformed into numerical representations using TF-IDF for machine learning models and word embeddings (GloVe and FastText) for deep learning models. Subsequently, multiple machine learning algorithms (Decision Tree, Random Forest, and Support Vector Machine) and deep learning architectures (CNN and BiLSTM) are trained and evaluated to capture both shallow and deep semantic patterns in the data.

To improve interpretability, Local Interpretable Model-agnostic Explanations (LIME) is applied to analyze model predictions and identify the most influential features contributing to classification decisions. This ensures transparency and increases trust in the proposed system.

Overall, the proposed framework integrates data preprocessing, augmentation, feature extraction, classification, and explainability into a unified pipeline for robust hate speech detection, where each stage is systematically connected to ensure effective learning and interpretation of the data. Figure [**?**] illustrates the complete methodology architecture, showing the flow and interconnection between all components of the proposed system.

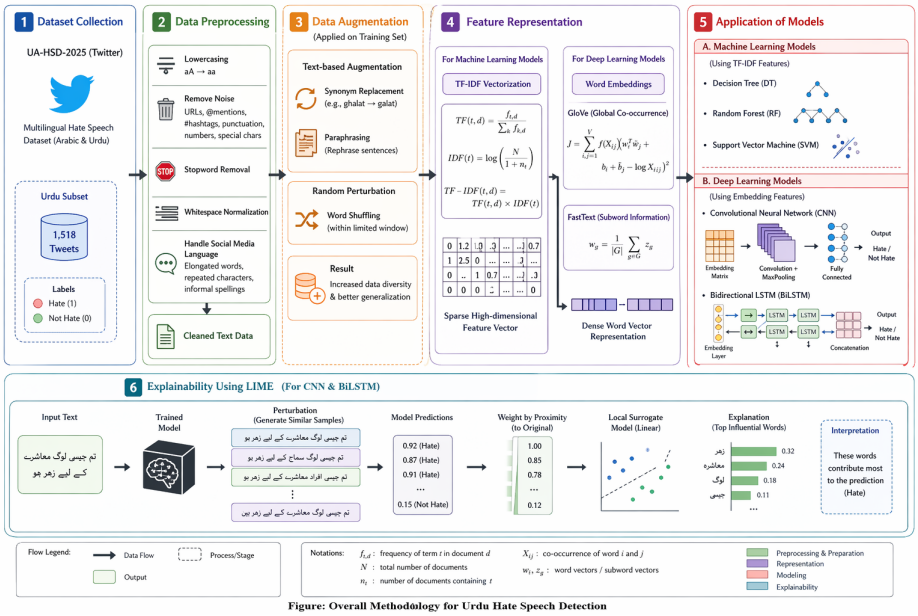


Figure: Overall Methodology for Urdu Hate Speech Detection

Fig. 1. Overall methodology architecture of the proposed hate speech detection framework.

6 Results and Analysis

In this section, we discuss the results based on the proposed methodology and evaluate the performance of both machine learning and deep learning models for hate speech detection in Urdu social media text. The experimental setup follows an 80–20 train-test split, where 80% of the data is used for training and 20% is used for testing to ensure fair evaluation of all models.

For machine learning models, standard configurations of Decision Tree, Random Forest, and Support Vector Machine are applied using TF-IDF features. For deep learning models, the experiments are conducted using the following parameters: maximum vocabulary size of 10,000, maximum sequence length of 100, embedding dimension of 300, 128 filters with a kernel size of 5, pooling size of 5, and 128 LSTM units. The models are trained for 9 epochs with a batch size of 16.

The results demonstrate that deep learning models outperform traditional machine learning approaches due to their ability to capture contextual and sequential information. Overall, this experimental setup provides a comprehensive comparison of different learning paradigms for effective hate speech detection.

Table 1. Performance comparison of machine learning models for binary hate speech detection.

Model	Precision	Recall	F1-score	Accuracy
SVM	0.975	0.974	0.973	0.974
RF	0.984	0.984	0.984	0.984
DT	0.974	0.974	0.974	0.974

6.1 Results for Machine Learning Models

Table 1 shows the performance comparison of machine learning models for binary hate speech detection, where the task is to classify text into *hate* and *not hate* categories. The evaluation is based on precision, recall, F1-score, and accuracy to ensure a comprehensive performance assessment.

The results indicate that all models perform strongly on the binary classification task. Among them, the Random Forest (RF) model achieves the best performance with an accuracy of 98.4%, along with consistently high precision, recall, and F1-score values of 0.984. This demonstrates the effectiveness of ensemble learning in capturing complex patterns within the TF-IDF feature space.

The Support Vector Machine (SVM) also shows competitive performance with an accuracy of 97.4%, highlighting its robustness in handling high-dimensional sparse textual representations. Similarly, the Decision Tree (DT) model achieves comparable results; however, its performance is slightly lower due to its sensitivity to data variations and tendency toward overfitting.

Overall, these findings suggest that ensemble-based approaches such as Random Forest provide better generalization for hate speech detection, while all models demonstrate strong capability in distinguishing between hate and non-hate content in Urdu social media text.

6.2 Results for Deep Learning Models

Table 2 presents the performance comparison of deep learning models for binary hate speech detection using different word embedding techniques, including FastText and GloVe. The evaluation is carried out using precision, recall, F1-score, and accuracy metrics to assess model effectiveness in distinguishing between *hate* and *not hate* classes.

The results show that BiLSTM with FastText embeddings achieves the best overall performance, reaching a remarkable accuracy of 99.3% along with equally high precision, recall, and F1-score values. This indicates that FastText effectively captures subword-level information, which is particularly useful for handling informal and morphologically rich Urdu text commonly found in social media.

In contrast, CNN with FastText embeddings also performs well, achieving an accuracy of 94.7%, demonstrating its ability to extract local semantic patterns from text. However, its performance is slightly lower than BiLSTM due to its limited capability in modeling long-term dependencies.

Table 2. Performance comparison of deep learning models using FastText and GloVe embeddings for binary hate speech detection.

Model	Precision	Recall	F1-score	Accuracy
FastText + CNN	0.947	0.947	0.946	0.947
FastText + BiLSTM	0.993	0.993	0.993	0.993
GloVe + CNN	0.899	0.508	0.460	0.508
GloVe + BiLSTM	0.840	0.799	0.713	0.799

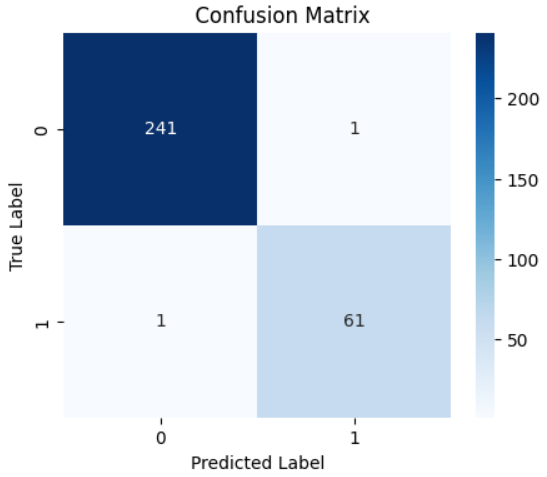


Fig. 2. Confusion matrix of the proposed hate speech detection model.

When using GloVe embeddings, a noticeable performance drop is observed. CNN achieves only 50.8% accuracy, while BiLSTM improves performance to 79.9%. This suggests that GloVe embeddings are less effective for noisy and context-dependent social media text compared to FastText, which benefits from subword information and better generalization.

Overall, the results demonstrate that BiLSTM combined with FastText embeddings provides the most robust performance for hate speech detection, highlighting the importance of both contextual modeling and effective word representation. Figure 2 illustrates the confusion matrix of the proposed model, which shows the distribution of correctly and incorrectly classified instances across the *hate* and *not hate* classes, thereby providing a detailed understanding of the model's classification performance.

6.3 Interpretability with LIME

The interpretability of the proposed model is demonstrated in Figure 3, where a sample Urdu text is analyzed using LIME. The original Urdu sentence shown in

یہ شیطانی سانپوں کا جھنڈ ہیں، ان کا سر قلم کر کے زندہ دفن کر دینا چاہیے۔

the figure is translated into English as ”*These people are extremely hateful and should not be trusted*”.

The proposed BiLSTM model with FastText embeddings classifies this instance as *Hate* with a confidence score of 1.00, indicating a highly confident prediction. The LIME explanation highlights the contribution of individual words toward the final decision, where words shown in blue strongly support the *Hate* class, while those in orange contribute toward the *Not Hate* class.

It can be observed that specific offensive and contextually negative terms in the Urdu text play a dominant role in driving the prediction toward the hate category. This demonstrates that the model effectively captures semantic meaning and contextual dependencies within the text.

Furthermore, the use of FastText embeddings enables the model to understand subword-level information, which is particularly beneficial for handling informal and morphologically rich Urdu language. Overall, the figure confirms that the proposed model is not only highly accurate but also interpretable, as it provides clear insights into which linguistic features influence the classification decision.

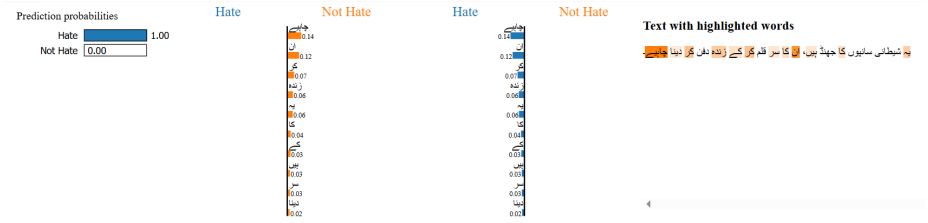


Fig. 3. LIME-based explanation of the proposed BiLSTM model with FastText embeddings for Urdu hate speech classification.

7 Conclusion and Future Work

This study addressed the growing threat of hate speech on social media by proposing an explainable deep learning model for Urdu, a low-resource and morphologically rich language. Unlike traditional black-box approaches, our framework integrates a FastText-based BiLSTM model with LIME explanations, achieving 99.3% accuracy on the UA-HSD-2025 dataset while providing word-level interpretability. The results demonstrate that combining subword-level embeddings with bidirectional contextual modeling effectively captures informal and ambiguous hate speech patterns. By transforming the detection process from an opaque decision system into a transparent one, this work enhances user trust and facilitates real-world deployment in content moderation systems.

Despite these promising outcomes, several limitations remain. The model was evaluated on a single dataset, which may limit generalizability. Future work

should focus on cross-dataset and cross-platform validation to assess robustness. Additionally, the current framework relies on supervised learning with manually annotated data, which is costly and time-consuming for low-resource languages. Exploring semi-supervised or few-shot learning approaches could reduce annotation dependency. Another promising direction is integrating multilingual and code-mixed hate speech detection, as Urdu users frequently blend languages. Finally, extending explainability beyond LIME to counterfactual or concept-based explanations may further improve model transparency and user confidence in real-time moderation systems.

References

1. Abubakar, H.D., Umar, M., Bakale, M.A.: Sentiment classification: Review of text vectorization methods: Bag of words, tf-idf, word2vec and doc2vec. SLU Journal of Science and Technology 4(1), 27–33 (2022)
2. Ahmad, M., Waqas, M., Hamza, A., Usman, S., Batyrshin, I., Sidorov, G.: Ua-hsd-2025: Multi-lingual hate speech detection from tweets using pre-trained transformers. Computers 14(6), 239 (2025)
3. Aksoy, Ç., Demirezen, M.U., Sağıroğlu, Ş.: Hate speech detection in turkish: An ensemble transformer-based deep learning approach. Engineering Applications of Artificial Intelligence 164, 113147 (2026)
4. Albladi, A., Islam, M., Das, A., Bigonah, M., Zhang, Z., Jamshidi, F., Seals, C.: Hate speech detection using large language models: A comprehensive review. IEEE Access 13, 20871–20892 (2025)
5. Alemerien, K., Al-Ghareeb, A., Alksasbeh, M.Z.: Sentiment analysis of online reviews: A machine learning based approach with tf-idf vectorization. Journal of Mobile Multimedia 20(5), 1089–1116 (2024)
6. Castaño-Pulgarín, S.A., Suárez-Betancur, N., Vega, L.M.T., López, H.M.H.: Internet, social media and online hate speech: Systematic review. Aggression and Violent Behavior 58, 101608 (2021)
7. Chapagain, S., Hamdi, S.M.: Advancing hate speech detection with transformers: Insights. In: Advances in Social Networks Analysis and Mining: Proceedings of the 17th International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2025). p. 432. Springer Nature (2025)
8. Chetty, N., Alathur, S.: Hate speech review in the context of online social networks. Aggression and Violent Behavior 40, 108–118 (2018)
9. Costa Ribeiro, L., Paes, A.: Does f1 fool you? a survey on annotator metadata transparency in hate speech detection. Journal of Information, Communication and Ethics in Society pp. 1–21 (2026)
10. Eilertsen, B., Björgfinsdóttir, R., Vargas, F., Ramezani-Kebrya, A.: Aligning attention with human rationales for self-explaining hate speech detection. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 37369–37378. AAAI Press (March 2026)
11. Geetanjali, Kumar, M.: Exploring hate speech detection: Challenges, resources, current research and future directions. Multimedia Tools and Applications 84(31), 38423–38459 (2025)
12. Kaur, S., Singh, S., Kaushal, S.: Abusive content detection in online user-generated data: A survey. Procedia Computer Science 189, 274–281 (2021)

13. Kiritchenko, S., Nejadgholi, I., Fraser, K.C.: Confronting abusive language online: A survey from the ethical and human rights perspective. Journal of Artificial Intelligence Research 71, 431–478 (2021)
14. Mishra, S., Thakur, S., Mamidi, R.: Enhancing hate speech detection on social media: A comparative analysis of machine learning models and text transformation approaches. arXiv preprint arXiv:2602.20634 (2026)
15. Mukherjee, S., Das, S.: Application of transformer-based language models to detect hate speech in social media. Journal of Computational and Cognitive Engineering 2(4), 278–286 (2023)
16. Qureshi, M.D.M., Qureshi, M.A., Rashwan, W.: Explainable ai for hate speech moderation: A stakeholder-centered and sociotechnical review. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 16(1), e70076 (2026)
17. Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A., Silva, C.: A comprehensive review on automatic hate speech detection in the age of the transformer. Social Network Analysis and Mining 14(1), 204 (2024)
18. Rawat, A., Kumar, S., Samant, S.S.: Hate speech detection in social media: Techniques, recent trends, and future challenges. Wiley Interdisciplinary Reviews: Computational Statistics 16(2), e1648 (2024)
19. Rogers, D., Preece, A., Innes, M., Spasić, I.: Real-time text classification of user-generated content on social media: Systematic review. IEEE Transactions on Computational Social Systems 9(4), 1154–1166 (2021)
20. Umansky, N., Kubli, M., Kotarcic, A., Bronner, L., Kurer, S., Grech, P., Donnay, K.: Improving hate speech detection with large language models. European Journal of Political Research 1, 12 (2026)
21. Usman, M., Ahmad, M., Sidorov, G., Gelbukh, A., Tellez, R.Q.: A large language model-based approach for multilingual hate speech detection on social media. Computers 14(7), 279 (2025)
22. Xue, Q., Dou, Y., Shi, Z.R., Li, X., Gao, W.: Mmbert: Scaled mixture-of-experts multimodal bert for robust chinese hate speech detection under cloaking perturbations. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 34196–34204. AAAI Press (March 2026)
23. Zhang, L.: Feature extraction based on naive bayes algorithm and tf-idf for news classification. PLoS One 20(7), e0327347 (2025)

Electronic edition
Available online: http://www.rcs.cic.ipn.mx

http://rcs.cic.ipn.mx

Centro de Investigación
en Computación